Text Hierarchical Clustering

07 February 2024 12:27

Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. It can be divided into three main levels:

1. Agglomerative Hierarchical Clustering: This approach begins with each data point as a separate cluster and then merges the closest pairs of clusters until all the data points belong to a single cluster.
2. Divisive Hierarchical Clustering: This approach starts with all data points in one cluster and then recursively divides the data into smaller clusters until each data point is in its own individual cluster.
3. Dendrogram: Hierarchical clustering results are often visualized using a dendrogram, a tree-like diagram that shows the arrangement of the clusters. The y-axis represents the distance or dissimilarity between clusters, and the x-axis represents the data points or clusters themselves.

Agglomerative Hierarchical Clustering:

This is the most common method of hierarchical clustering. It involves the following steps:

- Initialization: Start with each data point as its own cluster.
- Calculate Distance: Compute the distance (or dissimilarity) between each pair of clusters. Various distance metrics can be used, such as Euclidean distance, Manhattan distance, etc.
- Merge Closest Clusters: Identify the closest pair of clusters based on the distance metric and merge them into a single cluster.
- Update Distance Matrix: Recalculate the distances between the new cluster and all other clusters.
- Repeat: Repeat steps 2-4 until only a single cluster remains.
  Mathematically, this process can be represented as follows: - Let □□□dijrepresent the distance between cluster □$i$and cluster □j. - Define a linkage criterion (e.g., single linkage, complete linkage, average linkage) to determine the distance between clusters. - Update the distance matrix based on this criterion and continue merging until convergence.

Divisive Hierarchical Clustering:

Divisive hierarchical clustering is the opposite of agglomerative clustering. Instead of starting with individual data points and merging them into clusters, divisive clustering starts with all data points in a single cluster and then splits them recursively.

- Initialization: Start with all data points in one cluster.
- Calculate Separation: Measure the separation of the data points within the cluster.
- Divide Cluster: Split the cluster into two subclusters.
- Recursively Divide: Repeat steps 2-3 on each subcluster until a stopping criterion is met.
  Mathematically, divisive clustering involves determining the optimal split of the data at each step. This could be done using various methods, such as minimizing within-cluster variance or maximizing between-cluster variance.

Dendrogram:

A dendrogram is a graphical representation of the hierarchical clustering process. It displays the arrangement of clusters and the distances between them.

- Construction:
  - In an agglomerative approach, the dendrogram is built bottom-up, starting with individual data points as separate clusters and merging them step by step.
  - In a divisive approach, the dendrogram is built top-down, starting with all data points in a single cluster and recursively dividing them.
- Interpreting a Dendrogram:
  - The vertical lines in a dendrogram represent clusters.
  - The height of each vertical line represents the distance (or dissimilarity) between the clusters being merged.
  - The order in which clusters are merged can be inferred from the dendrogram's structure.
- Cutting the Dendrogram:
  - A horizontal line can be drawn across the dendrogram to determine the number of clusters desired.
  - Each cluster is formed by cutting the dendrogram at a particular height.

WARDS DISTANCE Example

X (Dont Read)

Example:

Imagine you have three clusters: □={□1,□2}A={a1,a2}, □={□1,□2}B={b1,b2}, and □={□1}C={c1}, with the following 2D points:

- □1=(1,1)a1=(1,1), □2=(2,2)a2=(2,2)
- □1=(5,5)b1=(5,5), □2=(6,6)b2=(6,6)
- □1=(0,5)c1=(0,5)

Let's go through the process of merging these clusters using Ward's method:

1. Calculate Cluster Centroids:
   - □A's centroid: (1.5,1.5)(1.5,1.5)
   - □B's centroid: (5.5,5.5)(5.5,5.5)
   - □$C$is a single point, so its centroid is (0,5)(0,5)
2. Calculate the Initial SS for Each Cluster:
   - □□(□)$SS(A)$: Sum of squared distances of □1a1and □2a2to the centroid of □A.
   - □□(□)$SS(B)$: Sum of squared distances of □1b1and □2b2to the centroid of □B.
   - □□(□)$SS(C)$: Since □Chas only one point, □□(□)=0$SS(C)$=0.
3. Calculate the Increase in SS for All Possible Merges:
   - Δ□□□□∪□Δ$SSA∪B$, Δ□□□□∪□Δ$SSA∪C$, and Δ□□□□∪□Δ$SSB∪C$need to be calculated to determine the increase in variance for each possible merge.
4. Determine the Merge with the Minimum Δ□□ΔSS:
   - Suppose Δ□□□□∪□Δ$SSA∪C$is the smallest. Then, clusters □$A$and □Cwould be merged.
5. Update Clusters and Repeat:
   - Now you have two clusters: □∪□$A∪C$and □B, and you would repeat the process until all points are in a single cluster.

Linkage Criteria

After merging clusters in agglomerative hierarchical clustering, the distance between the newly formed cluster and other clusters needs to be recalculated. There are different methods to calculate this distance, which are often referred to as linkage criteria. Three common linkage criteria are:

1. Single Linkage (or Minimum Linkage): This method calculates the distance between the closest points in the two clusters being merged. It tends to produce elongated clusters.
2. Complete Linkage (or Maximum Linkage): This method calculates the distance between the farthest points in the two clusters being merged. It tends to produce more compact, spherical clusters.
3. Average Linkage: This method calculates the average distance between all pairs of points in the two clusters being merged. It provides a balance between single and complete linkage.

Ques. When two clusters are merges say, A and b , how are the cluster cordinates defined

Ans.

To calculate the distance between a cluster (such as {A, B}) and a single point (like E) or between two clusters (such as {A, B} and {C, D}), we need to choose a linkage criterion. The linkage criterion determines how the distance between clusters is defined. Here are the common linkage criteria and how they would apply to calculating these distances:

1. Single Linkage (Nearest Point)

- Distance between {A, B} and E: Calculate the distance from E to both A and B. The distance between the cluster {A, B} and point E is the minimum of these distances. Mathematically, it's min(distance(A, E), distance(B, E)).
- Distance between {A, B} and {C, D}: Calculate all pairwise distances between points in the first cluster (A and B) and points in the second cluster (C and D). The distance between the two clusters is the minimum of these distances. Mathematically, it's min(distance(A, C), distance(A, D), distance(B, C), distance(B, D)).

2. Complete Linkage (Farthest Point)

- Distance between {A, B} and E: Calculate the distance from E to both A and B. The distance between the cluster {A, B} and point E is the maximum of these distances. Mathematically, it's max(distance(A, E), distance(B, E)).
- Distance between {A, B} and {C, D}: Calculate all pairwise distances between points in the first cluster (A and B) and points in the second cluster (C and D). The distance between the two clusters is the maximum of these distances. Mathematically, it's max(distance(A, C), distance(A, D), distance(B, C), distance(B, D)).

3. Average Linkage

- Distance between {A, B} and E: Since E is a single point, you can average the distances from E to A and B. The distance is (distance(A, E) + distance(B, E)) / 2.
- Distance between {A, B} and {C, D}: Calculate all pairwise distances between points in the first cluster (A and B) and points in the second cluster (C and D), then take the average of these distances. Mathematically, it's (distance(A, C) + distance(A, D) + distance(B, C) + distance(B, D)) / 4.

4. Ward's Method

- Ward's method is a bit more complex because it is based on minimizing the variance within the clusters. When merging two clusters, the choice is made so that it results in the minimum increase in total within-cluster variance after the merge. This requires calculating the sum of squared differences within all clusters and can be more computationally intensive than the other methods.

For the actual calculations, if the points are in a Euclidean space, the Euclidean distance formula is typically used:

Distance=(□2−□1)2+(□2−□1)2Distance=$(x2-x1)2+(y2-y1)2$

for points in a 2-dimensional space, or its equivalent in higher dimensions. If the data points have categorical attributes, other distance measures like the Hamming distance might be more appropriate.